The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 403-414

Testing Homogeneity for Random Effects in Linear Mixed Model

Chul H. Ahn[1)]

Abstract

A diagnostic tool for testing homogeneity for random effects is proposed in unbalanced linear mixed model based on score statistic. The finite sample behavior of the test statistic is examined using Monte Carlo experiments to examine the chi-square approximation of the test statistic under the null hypothesis.

Keywords : score test, maximum likelihood, heteroscedasticity

1. 서론

선형모형에서 가장 기본적인 가정은 등분산 (homoscedasticity) 즉, 오차의 분산이 설명변수의 모든 수준에서 동일하다는 것이다. 선형회귀나 고정효과를 갖는 선형모델에서는 이러한 등분산과 관련된 진단문제가 매우 활발히 논의되어 왔었다. 이러한 예는 Anscombe (1961), Bickel (1978), Hammerstrom (1981), Carrol 과 Ruppert (1981) 그리고 Cook 과 Weisberg (1983) 등에서 쉽게 찾아 볼 수 있으며, Atkinson (1985) 과 Cook 과 Weisberg (1982)에서는 이와 관련된 연구 및 연구결과들을 잘 정리해 놓고 있다.

선형혼합모형(linear mixed-effects model 또는 간단히 linear mixed model)에서 임의효과(random effects)와 오차의 분산과 관련된 논문들을 살펴보면, Hocking(1984)은 분산의 추정치가 얼마나 정확한지를 찾아내는 방법을 제시하고 이 추정치가 음수가 되는 이유를 알아내기 위한 그림을 유도하였다. Cook, Beckman 과 Nachtsheim (1987)은 Cook (1986)이 발표한 local influence 방법을 이용하여 오차분산, 오차공분산, 그리고 임의효과의 분산에 대한 perturbation 효과를 알아내는 방법을 제시하였다. 최근에는 Verbeke & Lasaffre (1996)가 선형혼합모형에서 임의효과에 대한 정규성 가정이 임의효과의 추정치에 미치는 영향을 조사하였다. 그들은 임의효과의 실제 분포가 몇개의 정규분포들의 혼합으로 이루어질 경우에 실제와는 다른 정규성이 가정된다면 임의효과의 추정치는 매우 부정확할 수 있다는 것을 밝혔다.

이 논문에서는 선형혼합모형(linear mixed model)에 있어서 임의효과에 대한 등분산 검정문제를

1) Associate Professor, Department of Applied Mathematics, Sejong University, 98 Kunja-dong, Kwangjin-ku, Seoul 143-747
E-mail : chahn@kunja.sejong.ac.kr

다룰 것이다. Ahn (1990)에서는 균형자료에 대한 등분산 문제를 다루었으나 여기에서는 불균형자료에 대한 등분산 문제를 다룰 것이며 균형자료는 불균형자료의 특수한 경우로 제시될 것이다. 임의효과가 등분산을 갖지 못할 때 임의효과의 분산은 설명변수들중 하나 또는 그 이상의 설명변수들과 함수관계를 갖거나 또는 시간이나 공간과 같은 변수들과도 함수관계를 가질 수 있다. 이 논문에서는 이러한 함수관계를 모형화하고 등분산을 진단하기 위한 검정법을 개발하게 될 것이다. 이를 위해 Silvey (1959)가 Lagrangian Multiplier Test 로 제시했던 스코어검정법을 사용할 것이다.

통계학의 여러 분야에서 등분산과 관련된 문제를 해결하기 위해 스코어검정이 사용되었다. 선형회귀분야에서 Cook 과Weisberg (1983)는 오차분산의 로그를 설명변수들의 선형함수로 표현하고 이를 이용하여 오차분산의 등분산 여부를 검정하는 스코어검정통계량을 개발하였다. Longitudinal data 분야에서는 Chi 와 Reinsel (1989)이 조건부독립 임의효과모형(conditional independence random effects model)에 있어서 그룹내 오차들 사이에 자기상관이 존재하는지 여부를 알아내기 위한 스코어검정통계량을 개발하였다. 일반화 선형모형(generalized linear model)에서는 Smyth (1989)가 분산에 대해 로그링크(log-link)를 갖는 감마일반화 선형모형(gamma generalized linear model)을 가정하고 등분산여부를 찾아내는 스코어검정법을 제시하였다.

이 논문은 5개의 장으로 구성된다. 이 논문의 2장에서는 임의효과의 분산에 대한 몇가지 확장모형을 수립하고 3장에서는 이를 위한 기반으로 하여 등분산검정을 위한 스코어검정통계량을 유도한다. 4장에서는 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사의 타당성을 알아보기 위해 균형자료와 불균형자료에 대해서 간단한 시뮬레이션이 시행된다. 마지막 5장에서는 여기에서 제시된 스코어검정통계량의 유효성을 알아보기 위한 좀더 광범위한 시뮬레이션의 필요성과 추가 확장모델 등 추후 연구 대상에 대한 논의와 결론을 내린다.

2. 모델 확장

선형혼합모형은 다음과 같이 쓰여 질 수 있다.

$$y = X\beta + Ub + \varepsilon \tag{1}$$

여기서, y 는 N x 1 의 반응벡터이며 y_{ij} 를 각각의 구성요소로 갖는다. y_{ij} 는 i 번째 그룹의 j 번째 관측값을 나타낸다. 총 그룹수는 t 이고 i 번째 그룹의 관측값 갯수는 n_i 이다. X 는 N x p 행렬로서 그 구성요소는 이미 아는 값들이다. β 는 p x 1 벡터이고, 절편을 포함하고 있을 경우는 (p+1) x 1 벡터로서 미지의 회귀모수이고 고정효과이다. U 는 N x t 행렬로서 임의효과를 계획(design)하기 위한 것이고, b 는 t x 1 벡터로서 서로 독립적인 요소인 b_i 를 갖고 있는데, b_i 는 평균 0, 분산 σ_b^2 을 갖는 정규분포를 따르는 것으로 가정하고 있다. 오차벡터인 ε는 서로 독립적이며 구성요소로서 ε_{ij} 를 갖고, 그 각각은 평균 0, 분산 σ^2 을 갖는 정규분포를 따르는 것으로 가정된다. 확률벡터인 b 와 ε 는 독립으로 가정된다. 모형 (1)은 오직 하나의 임의효

과 만을 포함하고 있으나, Ub 를 $\Sigma_{i=1}^{c} U_i b_i$ 로 대치하는 경우, 둘 이상의 임의효과를 포함시킬 수 있다. U 는 t 개의 벡터, 즉 $(u_1, u_2, \cdots, u_t)$ 로 구성된다. 여기서 u_i는 N x 1 벡터로서 그룹 i 에 해당하는 자리에는 1을, 그외의 자리에는 0을 갖는다. U 가 설명변수를 나타내는 벡터를 갖는 경우, 모형 (1)은 임의계수모형(random coefficient model)을 포함하게 된다.

선형혼합모형에서 한가지 중요한 가정은 각각의 임의효과들이 똑같은 정규모집단에서 나왔다고 하는 것이다. 이제부터 이 가정의 유효함을 알아보기 위한 진단방법을 찾아 보기로 하자. 한가지 방법은 모델확장을 통하는 것이다. 즉, 모형 (1)에 모수를 추가하여 확장된 모형을 만들고 이 추가된 모수에 대해 스코어검정을 실시하는 것이다. 모형 (1)의 임의효과에 대해 아래와 같은 모델확장을 생각해 볼 수 있다.

$$var(b_i) = \sigma_b^2 h(\lambda^t z_i) \tag{1.1}$$

여기서 z_i 는 그룹 i 에 대한 q x 1 공변수(covariate)벡터이고, 구성요소는 z_{ik} 로서 아래인자 k 는 1 에서 q (q 는 공변수의 갯수)까지의 값을 취한다. λ 는 q x 1 벡터의 미지 모수이다. 함수 $h(\cdot)$ 는 λ 에 대해 두번 미분 가능한 함수이며 모든 z_i 에 대해 $h(\lambda_0{}^t z_i) = 1$ 를 만족하는 특정한 값인 λ_0 가 존재한다고 가정한다. 따라서 등분산을 위한 검정은 자연히 $\lambda = \lambda_0$ 가 될 것이다. 함수 h 가 취할 수 있는 형태중 가장 유용한 것은 지수함수 일 것이다. 왜냐하면, 지수함수는 계속 미분 가능하고 $\lambda = 0$ 일때 등분산을 회복하게 되기 때문이다. 뿐만이 아니라 지수는 단조함수(monotonic function)이므로 분산이 어떤 방향으로 증가하는 형태를 보일 경우, 그 방향을 쉽게 찾을 수 있다는 것이다. 즉, 분산이 증가하는 방향은 $\lambda^t z_i$ 일 것이다. 첫번째로 생각할 수 있는 함수로는 다음과 같다.

$$h(\lambda^t z_i) = exp(\lambda^t z_i) \tag{2}$$

여기서 z_i 는 앞에서 정의되었듯이 그룹 i 를 위한 공변수 벡터로서 대개는 설명변수인 X 의 행렬에서 취해진다. 때때로 벡터 z_i 가 λ 에 곱해지기 전에 제곱이나 로그 변환을 취할때 이분산(heteroscedasticity)이 탐지되기도 한다. 따라서 식(2)에서 한 걸음 더 나아가 다음과 같은 함수를 생각할 수 있다.

$$h(\lambda^t z_i) = exp(\sum_{k=1}^{q} \lambda_k z_{ik}^{\alpha_k}) \tag{3}$$

식 (3)에서 $\alpha_k = 0$ 은 로그변환을 가리킨다. 분산이 반응변수의 기대값에 따라 달라질 때에는 h 가 아래의 식 (4) 와 같은 함수형태를 갖는 것이 유용할 것이다.

$$h(\lambda^t z_i) = h(E(y_i)) \tag{4}$$

식(1.1)에 있는 확장모델이 암시하고 있는 것은 그룹효과인 b_i 가 하나의 정규모집단에서 나왔다고 하기에는 값들이 너무 분산되어 있다는 것이며, 이러한 분산의 크기는 z_i 를 통하여 모형화 할 수 있다는 것이다.

3. 스코어검정

확률밀도함수 $f(y;\theta)$ 를 갖는 확률벡터 y 를 생각해보자. θ 는 $\theta\varepsilon\Theta\subseteq R^r$로 표시되는 r x 1 모수벡터이다. 확률밀도함수 $f(y;\theta)$ 는 Serfling (1980, p. 144)의 정규조건(regularity conditions)을 만족한다고 가정한다. 확률벡터 y 에서 얻어지는 t 개의 독립적인 관측치벡터는 $y_1, y_2, \cdots, y_t$ 로 표시하기로 하자. 그러면, 확률벡터 $y_1, y_2, \cdots, y_t$ 들의 로그우도함수인 $l(\theta)$ 는 $l(\theta)= l_1(\theta)+ l_2(\theta)+\cdots+ l_t(\theta)$ 로 표시된다. 여기서, $l_i(\theta)=\log f(y_i;\theta)$ 이다. 이제 θ 가 θ_1 과 θ_2 로 분할된다고 하자. 즉, $\theta= (\theta_1^t, \theta_2^t)^t$. 그리고, θ_2 는 q x 1 모수벡터라고 하자. 이제, 스코어벡터와 정보행렬의 구성요소는 각각 다음과 같이 쓰여진다.

$$d_j=\sum_i \partial l_i(\theta)/\partial\theta_j\ ,\quad J_{jk}=-E[\sum_i \partial^2 l_i(\theta)/\partial\theta_j\,\partial\theta_k^t].$$

여기서 $\sum$ 의 인자 i 는 1 에서 t 까지의 값을 취하고, 인자 j와 k 는 1 과 2 를 취한다. $\hat{d}_j$과 $\hat{J}_{jk}$ 를 각각 d_j 과 J_{jk} 이 $\theta_1=\hat{\theta}_1$ ($\theta_2=0$에서 최우추정량) 일 때 평가된 통계량이라 하자. 이제, 귀무가설, H_0: $\theta_2=0$ 와 대립가설, H_1: $\theta_2\neq 0$ 을 검정하는 스코어검정통계량은 Cox 과 Hinkley (1974, 9장)에 의하면 다음과 같이 쓰여질 수 있다.

$$S=\hat{d}_2^{\,t}(\hat{J}_{22}-\hat{J}_{21}\hat{J}_{11}^{-1}\hat{J}_{12})^{-1}\hat{d}_2 \tag{5}$$

식 (5)의 스코어검정통계량 S 는 귀무가설이 H_0: $\theta_2= 0$ 일때 점근적 (asymptotically)으로 자유도 q 를 갖는 χ^2 분포를 따른다. 스코어검정은 점근적으로 우도비검정(likelihood ratio test)과 같다. 다만, 스코어검정은 귀무가설에서만 추정이 이루어지면 검정통계량이 계산되었으나, 우도비검정은 대립가설에서도 추정이 이루어져야 한다는 점에서 종종 스코어검정이 선호되기도 한다. 이제, 분산에 대한 확장모형 (1.1)을 생각해 보자. 그룹 i 의 반응벡터 y_i 가 갖는 분산-공분산행렬은 다음과 같이 쓰여질 수 있다.

$$Cov(y_i) = \sigma^2 I_i+ \sigma_b^2 h(\lambda_t z_i) 1_i\, 1_i^t \tag{6}$$

식 (6)에서 y_i 는 n_i x 1 벡터이며, I_i 는 n_i x n_i 단위행렬이고, 1_i 는 n_i x 1 의 열행렬 (column vector) 로서 모두 1 의 값을 갖는다. 이제 ξ 를 두 분산, ${\sigma_b}^2$ 와 σ^2 의 비율로 표시하면 분산-공분산행렬, $Cov(y_i)$ 은 다음과 같이 간단히 쓰여진다. 즉, $Cov(y_i)=\sigma^2 Q_i$ 이고, 여기서 $Q_i = I_i + \xi h(\lambda^t z_i) 1_i {1_i}^t$. 그리고 우도함수는 다음과 같이 주어진다.

$$l_i(\theta,\lambda) = -\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log|\sigma^2 Q_i| - \frac{1}{2}\,{\varepsilon_i}^t(\sigma^2 Q_i)^{-1}\varepsilon_i \quad (7)$$

여기서 $\theta=(\beta^t, \sigma^2, \xi)^t$, 그리고 $\varepsilon_i = y_i - {X_i}^t\beta$. 이제, 귀무가설, $H_0:\lambda=0$ 와 대립가설, $H_1:\lambda\neq 0$ 을 검정하기 위한 스코어검정통계량은 다음과 같다.

$$S = \hat{d}_\lambda{}^t(\hat{J}_{\lambda\lambda} - \hat{J}_{\lambda\theta}\hat{J}_{\theta\theta}^{-1}\hat{J}_{\theta\lambda})^{-1}\hat{d}_\lambda \ . \quad (8)$$

여기서 d_λ, $J_{\lambda\theta}$, $J_{\lambda\theta}$, $J_{\theta\theta}$, $J_{\theta\lambda}$ 는 각각 $d_\lambda = \partial l/\partial\lambda = \sum_i \partial l_i(\theta,\lambda)/\partial\lambda$,

$J_{\lambda\lambda} = -E[\sum_i \partial^2 l_i/\partial\lambda\partial\lambda^t]$, $J_{\lambda\theta} = -E[\sum_i \partial^2 l_i/\partial\lambda\,\partial\theta^t]$, $J_{\theta\theta} = -E[\sum_i \partial^2 l_i/\partial\theta\,\partial\theta^t]$ 그리고 $J_{\theta\lambda} = -E[\sum_i \partial^2 l_i/\partial\theta\,\partial\lambda^t]$ 로 쓰여지며, (8)에 쓰인 것은 이 식들에 MLE를 대입한 추정량이다. 이제 (8)에 주어진 스코어검정통계량을 풀면, 다음과 같은 결과를 얻는다.

(결과 1)

V 는 t x 1 확률벡터로서 다음과 같은 구성요소들을 갖는다고 하자.

$$v_i = {\phi_i}^2\,{\bar{e}_i}^2/\hat{\sigma}^2 - \phi_i \ , \quad (9)$$

식 (9)에서 $\phi_i = n_i/(1+n_i\hat{\xi})$, $\bar{e}_i = \sum_{j=1}^{n_i}(y_{ij} - {x_{ij}}^t\hat{\beta})/n_i$, 그리고 $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\xi}$ 은 귀무가설 아래에서 β, σ^2, ξ, 각각의 최우추정량이다. $h'(\lambda^t z_i) = \partial h(\lambda^t z_i)/\partial\lambda$ 라고 놓자. $h'(\lambda^t z_i)$ 는 $\lambda = \lambda_0$ 일때 계산된 값이다. C 는 t x q 행렬이고 이 행렬의 i 번째 행은 $[h'({\lambda_0}^t z_i)]^t$ 이다. C_ϕ 는 t x q 행렬이고 i 번째 행은 $\phi_i[h'({\lambda_0}^t z_i)]^t$ 이다. $\bar{C}$ 는 t x q행렬로서 C 의 각 열에서 그 열의 평균을 뺀 값이다. 마지막으로, F^t 와 B 는 각각 q x 2 그리고 2 x 2 행렬이고 다음과 같다.

$$F^t = \left(\sum_i \frac{\phi_i}{\hat{\sigma}^2}h'({\lambda_0}^t z_i)\ \sum_i {\phi_i}^2 h'({\lambda_0}^t z_i)\right) \ , \quad B = \begin{bmatrix} N/\sigma^4 & \sum_i \phi_i/\hat{\sigma}^2 \\ \sum_i \phi_i/\hat{\sigma}^2 & \sum_i {\phi_i}^2 \end{bmatrix}$$

그러면, $H_0: \lambda = \lambda_0$ 을 검정하기 위한 스코어검정통계량은 다음과 같이 표현된다.

$$S = \frac{1}{2} V^t \overline{C} [C_\phi{}^t C_\phi - F^t B^{-1} F]^{-1} \overline{C}^t V \qquad (10)$$

(증명)

로그우도함수를 λ 에 대하여 편미분하면 $d_\lambda = \partial l / \partial \lambda = \sum_i \partial l_i(\theta, \lambda)/\partial \lambda$ 로 쓰여진다. 여기서 $l_i(\theta, \lambda)$ 의 형태는 (7)에서 주어져 있다. d_λ 의 k 번째 요소는

$$d_{\lambda_k} = \sum_i \frac{\partial l_i}{\partial \lambda_k} = -\frac{1}{2} \sum_i [TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}) + \frac{1}{\sigma^2} \varepsilon_i{}^t (- Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}) \varepsilon_i]$$

로 쓰여지는데, 이 식은 Rogers(1980) 에 있는 행렬미분에 대한 아래의 두 결과를 이용하여 얻은 것이다.

$$\frac{\partial}{\partial \lambda_k} \log | Q_i | = TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}), \text{ 그리고 } \frac{\partial Q_i^{-1}}{\partial \lambda^k} = - Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}.$$

여기서, $Q_i = I_i + \xi h(\lambda^t z_i) 1_i 1_i{}^t$ 이고 그 역행렬은 Graybill (1969) 에서와 같이 $Q_i^{-1} = I_i - \xi h(\lambda^t z_i) 1_i (1 + \xi h(\lambda^t z_i) 1_i{}^t 1_i)^{-1} 1_i{}^t$ 로 쓰여진다. $\partial Q_i / \partial \lambda_k$ 을 풀어보면 $\xi \partial h(\lambda^t z_i)/\partial \lambda_k 1_i 1_i{}^t$ 이 얻어진다. 이것들을 $\lambda = \lambda_0$ 일때 구해보면, $h(\lambda_0{}^t z_i) = 1$ 이므로,

$$TR(Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k}) = h_{ik}' \frac{n_i}{1 + n_i \xi} \quad \text{그리고} \quad \varepsilon_i{}^t (Q_i^{-1} \frac{\partial Q_i}{\partial \lambda_k} Q_i^{-1}) \varepsilon_i = h_{ik}' (\frac{n_i \overline{\varepsilon}_i}{1 + n_i \xi})^2$$

이 얻어진다. 여기서 h_{ik}' 는 $\partial h(\lambda^t z_i)/\partial \lambda_k$ 을 $\lambda = \lambda_0$ 로 놓고 구한 것이며, $\overline{\varepsilon_i}$ 는 i 그룹의 평균 즉, $\overline{\varepsilon_i} = \sum_{j=1}^{n_i} \varepsilon_{ij} / n_i$ 이다. 이제, d_λ 에서 β, σ^2, ξ 그리고 λ 대신 귀무가설 아래에서 구해진 최우추정량 $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\xi}$ 과 $\lambda = \lambda_0$ 을 대입하면 스코어벡터 $\hat{d}_\lambda$ 은 다음과 같이 표현된다.

$$\hat{d}_\lambda = -\frac{\hat{\xi}}{2} \sum_i (\phi_i - \frac{\phi_i{}^2 \overline{e}_i{}^2}{\hat{\sigma}^2}) h_i'$$

여기서, $\phi_i = n_i / (1 + n_i \hat{\xi})$ 이고, h_i' 는 q x 1 벡터로서 h'_{ik} 을 k 번째 요소로 갖는다. $\overline{h}'$ 을 q x 1 벡터로서 그요소가 모두 똑같이 h_{ik}' 의 평균값을 갖는다고 하자. 그러면 σ^2에 대한 최우추

정량의 관계식 즉, $\hat{\sigma}^2=\sum_i \phi_i{}^2\ \bar{e}_i{}^2/\sum_i \phi_i$, 을 통하여

$$\hat{d}_\lambda=\frac{\hat{\xi}}{2}\sum_i(h_i'-\bar{h}')v_i$$

이 얻어지고, $\sum_i v_i=0$ 을 이용하여 $\hat{d}_\lambda$ 이 다음과 같이 간결하게 표현된다.

$$\hat{d}_\lambda=\frac{1}{2}\hat{\xi}\,\overline{C}^t V \tag{11}$$

한편, $J_{\lambda\lambda}$ 의 km 번째 요소는 다음과 같이 쓰여진다.

$J_{\lambda_k\lambda_m}=-E[\sum_i \partial^2 l_i/\partial\lambda_k\,\partial\lambda_m]=\frac{1}{2}\sum_i TR(Q_i^{-1}\frac{\partial Q_i}{\partial\lambda_k}Q_i^{-1}\frac{\partial Q_i}{\partial\lambda_m})$. $\hat{J}_{\lambda\lambda}$ 을 구하기 위하여 귀무가설 아래에서 A_{ik} 와 A_{im} 을 다음과 같이 정의한다.

$$A_{ik}=\partial Q_i/\partial\lambda_k=\xi h_{ik}'\,1_i 1_i^t \qquad A_{im}=\partial Q_i/\partial\lambda_m=\xi h_{im}'\,1_i\,1_i{}^t$$

여기서, h'_{im} 는 앞에서 h'_{ik} 이 정의되었듯이, $h'_{im}=\partial h(\lambda^t z_i)/\partial\lambda_m$ 로 정의된다.

$J_{\lambda_k\lambda_m}$ 는 $\lambda=\lambda_0$ 일 때 A_{ik}, A_{im}, h_{ik}', h_{im}' 을 이용, 다음과 같이 쓰여진다.

$$\begin{aligned}\frac{1}{2}\sum_i[TR(Q_i^{-1}A_{ik}Q_i^{-1}A_{im})&=TR(A_{ik}A_{im})-\frac{\xi}{1+n_i\xi}TR(1_i{}^tA_{ik}A_{im}1_i)]\\&-\frac{\xi}{1+n_i\xi}TR(1_i{}^tA_{im}A_{ik}1_i)+\frac{\xi^2}{(1+n_i\xi)^2}TR(1_i{}^tA_{ik}1_i\,1_i{}^tA_m1_i)\\&=\frac{1}{2}\sum_i[\,n_i{}^2\,\xi^2\,h_{ik}'h_{im}'(1-\frac{n_i\xi}{1+n_i\xi})^2]\end{aligned}$$

결국, $\hat{J}_{\lambda\lambda}$ 은 위에서 정의된 C_ϕ 을 이용하여 다음과 같이 간단히 표현된다.

$$\hat{J}_{\lambda\lambda}=\frac{1}{2}\,\hat{\xi}^2\,C_\phi{}^tC_\phi \tag{12}$$

$J_{\lambda\theta}$ 은 세부분으로 나누어진다. $J_{\lambda\theta}=[\,J_{\lambda\beta}\ \ J_{\lambda\sigma^2}\ \ J_{\lambda\xi}]$. 여기서, $J_{\lambda\beta}=0$

$J_{\lambda_k\sigma^2}=-E[\sum_i\partial^2 l_i/\partial\lambda_k\partial\,\sigma^2]=\frac{1}{2\sigma^2}\sum_i TR(Q_i^{-1}\frac{\partial Q_i}{\partial\lambda_k})$, 그리고

$J_{\lambda_k\xi}=-E[\sum_i\partial^2 l_i/\partial\lambda_k\partial\xi]=\frac{1}{2}\sum_i TR(Q_i^{-1}\frac{\partial Q_i}{\partial\lambda_k}Q_i^{-1}\frac{\partial Q_i}{\partial\xi})$.

$J_{\theta\theta}$ 는 $J_{\beta\beta}$, $J_{\beta\sigma^2}$, $J_{\sigma^2\sigma^2}$, $J_{\sigma^2\xi}$, $J_{\xi\xi}$ 로 구성되며 다음 행렬로 쓸 수 있다.

$$J_{\theta\theta}=\begin{bmatrix} J_{\beta\beta} & J_{\beta\sigma^2} & J_{\beta\xi} \\ J_{\sigma^2\beta} & J_{\sigma^2\sigma^2} & J_{\sigma^2\xi} \\ J_{\xi\beta} & J_{\xi\sigma^2} & J_{\xi\xi} \end{bmatrix}$$

여기서, $J_{\beta\beta}=\sum_i X_i{}^t Q_i^{-1} X_i/\sigma^4$,

$$J_{\beta\sigma^2}=J_{\beta\xi}=0,$$

$$J_{\sigma^2\sigma^2}=N/2\sigma^4,$$

$$J_{\sigma^2\xi}=\frac{1}{2}\sum_i TR(Q_i^{-1}\frac{\partial Q_i}{\partial\xi}),$$

$$J_{\xi\xi}=\frac{1}{2}\sum_i TR(Q_i^{-1}\frac{\partial Q_i}{\partial\xi}Q_i^{-1}\frac{\partial Q_i}{\partial\xi}).$$

$\hat{J}_{\lambda\lambda}$ 를 구하는 똑같은 방법을 이용하여 $\hat{J}_{\lambda\theta}$ 와 $\hat{J}_{\theta\theta}$ 을 구하면 다음 결과를 얻는다.

$$\hat{J}_{\lambda\theta}\hat{J}_{\theta\theta}^{-1}\hat{J}_{\theta\lambda}=\frac{1}{2}\hat{\xi}^2 F^t B^{-1}F. \tag{13}$$

(11), (12), (13) 을 정리하면 (10) 의 스코어검정통계량을 구한다. (증명 끝)

그룹별 관측치수가 똑같은 균형자료 ($n_i=n$) 의 경우, 스코어검정통계량 S는 다음과 같이 쓰여진다.

$$S=\frac{1}{2}R^t\overline{C}[\overline{C}^t\overline{C}]^{-1}\overline{C}^tR$$

여기서 R 은 t x 1 벡터이고 구성요소는 $R_i=\phi\bar{e}_i^2/\hat{\sigma}^2$ 이며, ϕ 는 $\phi=n/(1+n\hat{\xi})$ 이다. 이 경우, S 는 새로 만들어진 모형, $R=\gamma_0 1+C\gamma+\varepsilon_R$ 에서 R 을 종속변수로 하고 C 를 독립변수로 하여 적합된 회귀모형에서 얻어진 회귀제곱합의 1/2 에 해당한다. 만일 (2)에서 언급된 가중함수(weight function) h 로 지수함수를 사용한다면, $a_k=1$ 일 때 $h'(\lambda_0^t z_i)=z_i$ 이고, $a_k=0$ 일때 $h'(\lambda_0^t z_i)=\log(z_i)$ 이 될 것이다.

균형자료에서 얻어진 이 결과는 선형회귀분야에서 Cook 과 Weisberg (1983)에 의해 얻어진 결과와 원칙적으로 동일한 것이다. 그들이 보인 것은 선형회귀모형 $y_i=\beta_0+x_i^t\beta+\varepsilon_i$ 에서 오차분산의 확장모형으로 $var(\varepsilon_i)=\sigma^2\exp(\lambda^t z_i)$ 가 사용될 때 $H_0:\lambda=0$ 을 검정하기 위한 스코어검정통계량은 e_i 가 $e_i=y_i-\hat{\beta}_0-x_i^t\hat{\beta}$ 으로 표현되고, $\hat{\beta}_0$, $\hat{\beta}$, $\hat{\sigma}^2$ 가 귀무가설 아래에서

β_0, β, σ_2 의 최우추정량일때 $e_i^2/\hat{\sigma}^2$ 을 종속변수로 하고 z_i 을 독립변수로 하여 적합한 회귀모형에서 얻어진 회귀제곱합의 1/2 에 해당한다는 것이다. 분산이 관측치의 기대값 즉, $E(y_i)$ 에 따라 변할때에는, $h(\lambda^t z_i) = h(\lambda x_i^t \beta)$ 이 되고, $h'(\lambda^t z_i)$ 은

$$[\partial h(\lambda x_i^t \beta)/\partial \lambda]_{\lambda=\lambda_0} = x_i^t \hat{\beta}$$

에 의해 대치될 수 있다. 이 경우 우리가 알 수 있는 것은 스코어검정통계량이 함수 h 의 형태와는 무관하게 계산된다는 것이다.

4. 시뮬레이션

스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사가 유한표본을 갖고 있을 경우에 얼마나 정확하고 타당성이 있는 방법인가 알아보기 위해 균형자료와 불균형자료에 대해서 간단한 시뮬레이션이 시행되었다. 그룹수 t 는 5, 10, 그리고 20 을 갖도록 하였고, 설명변수의 갯수 p 는 1 로 고정시켰다. 균형자료의 경우, 그룹별 관측치의 갯수 n 은 5 로 고정하였다. 불균형자료의 경우, 그룹별 관측치의 갯수 n_i 는 난수표의 5 번째 column을 골라 아래로 읽으면서 4, 5, 6 에 해당하는 숫자를 골라 20 개를 마련하였다. 골라진 20개의 난수는 5, 6, 5, 5, 4, 4, 6, 6, 4, 6, 6, 4, 4, 5, 5, 4, 6, 5, 4, 4 였다. t 가 5인 경우 이중 처음 5 숫자인 5, 6, 5, 5, 4 를 n_i 로 사용하였다. 공변수 z 는 항상 설명변수 x 와 같도록 하였다. 공변수의 개수를 q 로 표시하고 있으므로 $q = p = 1$ 이 되도록 하였다. 설명변수행렬(design matrix), X 로 사용할 행렬을 만들기 위해 표준정규분포에서 임의수(random number)를 발생시켜 100 x 2 의 행렬을 준비하였다. 모형이 절편을 갖도록 하기 위해 X 의 첫번째 열은 모두 1 을 갖도록 하였다. t 값에 따라 우리는 499 개의 반복적인 표본을 추출하게 되고 각 표본마다 스코어검정통계량을 계산하게 된다. 499 개의 통계량이 모여지면 이것을 가지고 스코어검정통계량의 분포를 얻게 되는데 이렇게 한개의 스코어검정통계량의 분포를 얻게 되는 절차를 1 개의 시뮬레이션이라 하자. 즉, 우리는 균형인 경우 3개, 그리고 불균형인 경우 3개의 시뮬레이션이 필요하게 되며, 따라서 총 6 개의 시뮬레이션을 하게 된다. 각 시뮬레이션에서 t 의 값이 결정된면 이 행렬 X 는 필요한 부분을 위에서 준비한 100 x 2 의 행렬중 위 부분을 선택하여 사용하고 각 시뮬레이션이 시행되는 동안에는 값이 변하지 않고 고정될 것이다. 예를 들어 그룹수가 20 인 경우를 살펴보자. 균형자료인 경우에는 그룹별 관측치의 갯수가 5 이므로 관측치의 총 갯수는 100 이므로 100 x 2 의 행렬 모든 부분이 필요하지만 불균형자료의 경우에는 그룹별 관측치 갯수가 5, 6, 5, 5, 4, 4, 6, 6, 4, 6, 6, 4, 4, 5, 5, 4, 6, 5, 4, 4 (총 98 개) 이므로, X 의 왼쪽위코너에서 98 x 2 의 행렬을 취하게 된다. 이제, 그룹수가 10 이고 균형자료인 경우에는 그룹별 관측치의 갯수가 5 이므로 관측치의 총 갯수는 50 이므로 X 의 위부분에서 50 x 2 의 행렬을 취하게 된다. 불균형자료의 경우에는 그룹별 관측치 갯수가 5, 6, 5, 5, 4, 4, 6, 6, 4, 6 (총 51개) 이므로, X 의 위부분에서 51 x 2 의 행렬을 취하게 된다. 반응변수 y_{ij} 는 다음 모형에서 발생되었다.

$$y_{ij} = x_{ij}^{t}\beta + b_i + \varepsilon_{ij},\ i=1,\ldots,t,\ j=1,\ldots,n.$$

고정효과인 회귀모수 β 는 모두 0 으로 놓았고, σ^2 과 σ_b^2 또한 모두 1 로 놓아졌다. 즉, b_i 와 ε_{ij} 는 모두 기대값이 0 이고 분산이 1 인 표준정규분포에서 임의로 추출되었다.

(표 1) 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사

%점	t = 5	t = 10	t = 20	χ^2
0.90	1.71 (1.80)	2.39 (2.52)	2.58 (2.31)	2.71
0.95	2.09 (2.49)	3.23 (3.37)	3.74 (3.29)	3.84
0.975	2.37 (2.95)	4.36 (5.00)	5.03 (4.91)	5.02
0.99	2.81 (3.20)	5.56 (6.11)	7.06 (6.43)	6.63

위 (표 1)은 설명변수 p 가 하나인 경우, t 값에 따라 행하여진 6 개의 시뮬레이션 결과이다. 네개씩 짝을 지어 세로로 쓰여진 값들은 시뮬레이션에서 얻은 스코어검정통계량의 표본분포에서 읽은 90%, 95%, 97.5%, 그리고 99%에 해당하는 점들이다. 괄호 안의 값은 불균형자료인 경우에 얻어진 결과이다. 예를 들면, t =5 이고 p = 1 일때의 (1.71, 2.09, 2.36, 2.81) 과 (1.80, 2.49, 2.95, 3.20) 은 각각 균형자료와 불균형자료에서 얻어진 스코어검정통계량의 표본분포에서 읽은 90%, 95%, 97.5%, 그리고 99%에 해당하는 점들이다. 비교대상이 되는 χ^2 분포 (자유도 1)의 % 포인트는 오른쪽 끝 열에 2.71, 3.84, 5.02, 6.63 이 세로로 기재되어 있다. (표 1)에서 보여지듯이 균형자료와 불균형자료 모두, 그룹수 (t) 가 증가할수록 스코어검정통계량의 % 포인트는 χ^2 분포의 해당하는 점에 매우 가까운 것을 보여주고 있으며, 이는 스코어통계량의 점근적인 행태를 뒷받침해 준다고 할 수 있겠다. 그러나, 스코어검정통계량의 % 포인트가 해당 χ^2 값에 비해 일반적으로 작은 것으로 나타나고 있는데 이는 스코어검정을 위해서 χ^2 를 사용할 경우 보수적인 입장을 취하는 것이라고 할 수 있다. 이는 χ^2 검정을 사용함으로써 실제로 귀무가설을 기각해야하는 경우보다 적게 기각하게 되기 때문이다.

5. 결론

이 논문에서 제시된 스코어검정은 선형혼합모형에서 임의효과에 대한 등분산 여부를 검정하는데 쓰일 수 있을 뿐만 아니라 등분산 귀무가설이 기각되는 경우 그 원인을 찾는데 도움을 줄 수 있다. 예를 들면, 분산이 어느 방향으로 증가할 때 이 방향을 찾을 수 있다는 것이다. 스코어검정의 또 다른 장점은 통계량을 계산하기가 매우 용이하다는 것이다. 즉, 귀무가설 아래에서의 최우추정량만으로 검정통계량이 계산되므로 기존의 통계패키지로도 쉽게 구할 수 있기 때문이다. 또

한, 점근적으로는 우도비검정과 같을지라도 대립가설 아래에서 추정량을 필요로 하는 우도비검정과 달리 귀무가설 아래에서의 최우추정량만으로 검정통계량이 계산되므로 종종 우도비검정보다 선호될 수 있다.

여기에서 제시된 스코어검정통계량과 관련하여 앞으로 이루어져야 할 연구대상이 몇 가지 있다. 먼저, 소표본을 갖고 있을 경우에 스코어검정통계량의 귀무가설 분포에 대한 χ^2 근사가 얼마나 정확한지 알아보기 위한 시뮬레이션이 좀더 광범위하게 이루어져야 한다. 이 논문에서 다루고 있는 불균형자료의 경우에는 불균형의 정도에 따라 결과가 달라질 수 있으므로 시뮬레이션이 3장에서 행하여진 것보다 좀더 다양하고도 광범위하게 실시되어야 한다. 스코어검정통계량과 우도비검정통계량과의 비교 또한 시뮬레이션을 통하여 효과적으로 이루어질 수 있다. 그리고, (1.1) 에서 주어진 모델확장 이외에, 오차분산에 대한 다음과 같은 모델확장을 생각할 수 있다.

$$var(\varepsilon_{ij}) = \sigma^2 h(\lambda^t z_{ij}).$$

이 모형은 각각의 실험단위나 관측단위 사이에 존재하는 분산이 그 단위의 고유한 공변수(covariate)에 따라 달라질 수 있다는 것이다. 이 확장모형은 longitudinal 자료의 분석에 효과적으로 쓰일 수 있다. 왜냐하면, longitudinal 자료의 경우, 값이 커짐에 따라 관측치의 분산도 커지는 경향이 있기 때문이다.

선형혼합모형에서 스코어검정이 유용하게 쓰일 수 있는 분야로는 임의효과에 대한 정규성을 진단하는 일이라고 할 수 있다. 최근에 Verbeke 와 Lasaffre (1996)가 지적하였듯이 선형혼합모형에서 임의효과에 대한 정규성 가정이 임의효과의 추정치에 미치는 영향은 매우 중요하기 때문이다.

참고문헌

[1] Ahn, C.H. (1990), "Diagnostics for Heteroscedasticity in Mixed Linear Model," Journal of Korean Statistical Society, Volume XIX, No. 2, 171-175.

[2] Atkinson, A.C. (1985), Plots, Transformations and Regression, Oxford : Oxford University Press.

[3] Bickel, P. (1978), "Using Residuals Robustly I: Tests for Heterocesdasticity, Non-linearity", Annals of Statististics, Vol. 6, 266-291.

[4] Carroll, R.J. 와 Ruppert, D. (1981), "On Robust Tests for Heteroscedasticity," Annals of Statististics, Vol. 9, 205-209.

[5] Chi, E.M. 과 Reinsel, G.C. (1989), "Models for Longitudinal Data With Random Effects and AR(1) Errors," Journal of the American Statistical Association, Vol. 84, 452-459.

[6] Cook, R.D. (1986), "Assessment of Local Influence (with discussion)," Journal of Royal Statistical Society, Series B, 48, 133-169.

[7] Cook, R.D., Beckman, R. 과 Nachtsheim, C. (1987), "Diagnostics for Mixed-Model

Analysis of Variance, Technometrics 29, 413-426.

[8] Cook, R.D. 과 Weisberg, S. (1982), Residuals and Influence in Regression, London: Chapman Hall.

[9] ------ (1983), "Diagnostics for Heteroscedasticity in Regression," Biome-trika, 70, 1-10.

[10] Cox, D.R. 과 Hinkley, D.V. (1974), Theoretical Statistics, London: Chapman Hall.

[11] Hammerstrom, T. (1981), "Asymptotically Optimal Tests for the Heteroscedasticity in the General Linear Model," Annals of Statististics, Vol. 9, 368-380.

[12] Hocking (1984), "Diagnostics Methods in Variance Component Estimation," Proceedings of International Biometrics Conference, Tokyo, Japan.

[13] Rogers, G.S. (1980), Matrix Derivatives, Marcel Dekker, New York.

[14] Serfling (1980), Approximation Theorems of Mathematical Statistics, New York: Wiley.

[15] Silvey, S.D. (1959), "The Lagrangian Multiplier Test," The Annals of Mathematical Statistics, 30, 389-407.

[16] Smyth, Gordon (1989), "Generalized Linear Models with Varying Dispersion," Journal of Royal Statistical Society, Series B, 51, 47-60.

[17] Verbeke 와 Lasaffre (1996). "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," Journal of the American Statistical Association, Vol. 91, 217-221.